Detection of Hypergiants in AS-Level Topology Using Machine Learning

Michiko HARAYAMA† and Takuro KUDOH‡
†Department of Electric Electronics and Informatics, Faculty of Engineering, Gifu University
harayama@gifu-u.ac.jp
†‡1-1 Yanagido, Gifu, 501-1193, Japan
‡Department of Intelligence Science and Engineering, Graduate School of Natural Science and Technology, Gifu University
z4525032@edu.gifu-u.ac.jp

***Abstract*—With the spread of cloud services over the past two decades, the number of content holders (CHs) on the Internet has grown. Some of them are huge multinational companies such as GAFAM (Google, Apple, Facebook, Amazon, and Microsoft). In addition, content delivery networks (CDNs) have grown significantly with the increase in traffic from CHs to users. Some of these CHs and CDNs generate traffic levels comparable to those of major internet service providers (ISPs) and are called hypergiants (HGs). The impact of HGs is as large as that of major ISPs, and they need to be watched closely because the failure of any one of them may affect the entire Internet. Although it is not easy to capture the growth of individual autonomous systems (ASes), the detection of unknown growing HGs will be useful for preventing communication failures and understanding the impact status of ASes on the Internet.**

Therefore, in this study we attempted to detect unknown HGs by using publicly available data and machine learning methods. First, we extracted ASes from Tier 1 to Tier 3 from the AS relationship data published by the Center for Applied Internet Data Analysis (CAIDA) and analyzed the features of the ASes and the AS-level topology as a complex network. Next, we found that the random forest machine learning method was suitable for classifying ASes by their features, so we trained the features of famous HGs and detected HGs by using random forest. As a result, currently growing CDs and CHs were detected as HGs.

Keywords—AS-level topology, Hypergiants, Machine learning, Random forest, Detection

I. INTRODUCTION

With the spread of COVID-19, the use of the Internet as an established information infrastructure is accelerating. Telecommuting and online meetings and events are being encouraged, and more and more people are engaging in social activities via the Internet. The Internet is made up of interconnected networks called autonomous systems (ASes) [1], which are networks owned by organizations and operated under a single policy. Growth in the number of ASes has been accelerating. The Internet Assigned Numbers Authority (IANA) assigns an AS number (ASN) to each AS to identify it. According to the ASN assignment status by IANA [2], the number of ASes may exceed 100,000 by the end of 2021. Of these ASes, the number of net-connected devices is expected grow from 18.4 billion in 2018 to 29.3 billion in 2023 [3]. The bloated autonomous Internet, though artificial, is growing like an autonomous organism.

However, large-scale AS failures can significantly affect social activities: on November 6, 2017, a routing leak at Level3 (now CenturyLink [4]), a major US internet service provider (ISP), resulted in the worldwide advertisement of border gateway protocol (BGP) routing information for customers and connections that should have been internal. As a result, a large amount of traffic flowed across the Internet, and many users were unable to communicate due to the resulting congestion [5]. In such cases, each AS is required to respond to the connection structure and traffic situation of the entire Internet. However, each AS can understand only its own situation and surroundings, such as BGP routing information advertised by neighboring ASes. Therefore, the Center for Applied Internet Data Analysis (CAIDA) [6] regularly collects, analyzes, and publishes data at multiple points on the Internet in order to study the structure of the Internet and troubleshoot the network. CAIDA's scamper, Looking Glass [7] servers, and RIPE Atlas [8] measure and publish transmission of information about paths on the Internet.

ISPs have played a major role in the Internet. With the proliferation of cloud services, the number of content holders (CHs) who provide and sell content has increased, and their services have expanded. With the increase in traffic from CHs to users, content delivery networks (CDNs) have also increased in number and scale. CDN companies provide communication services that reduce the load on the network and enable faster delivery of content by CHs. Some CHs and CDNs, called hypergiants[1] (HGs), generate traffic levels comparable to those of major ISPs. Most Internet users use Google (Alphabet), Apple, Facebook, Amazon, and Microsoft (GAFAM), and thus these companies significantly affect the global economy. These HGs also affect Internet communication; for example, on August 25, 2017, Google sent incorrect routing information, resulting in a massive communication failure [5]. It is also known that P2P link overflows associated with HGs can influence Internet communication [9]. In addition, it is known that AS-level topology is changing and flattening due to the emergence of large-scale CH and CDH [10,11,12]. In addition, there is a report that Internet traffic is changing due to the lockdowns caused by the COVID-19 pandemic [13], and consequently ASes will change routing policies and cause AS-level topology to change.

To improve user services by avoiding failures and optimizing routing paths, each ISP needs to keep a close eye on the HGs that affect the entire Internet. One clue about which HGs can do that is the AS ranking based on customer cone size (CCS) published by CAIDA, but CCS indicates the influence as an ISP and does not that as CHs or CDNs. Therefore, we

[1] Major ISPs are also sometimes considered HGs, but ISPs are excluded from HGs in this study.

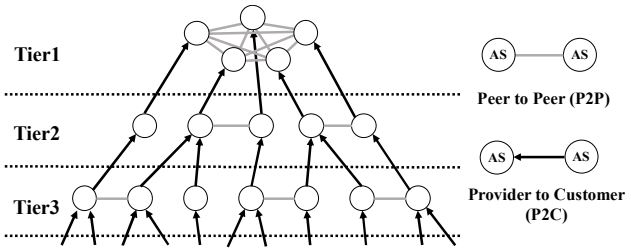


Fig. 1. AS-level topology

have conducted research to understand the structure of the Internet by analyzing CAIDA's public information on AS relationship data and BGP path lists of a stab AS [14-16]. By analyzing the AS-level topology and devising an index that reveals the influence of HGs, we tried to discover new emerging HGs. However, due to the complexity of factors influencing the structure of the Internet, it may be better to consider many features comprehensively than to identify HGs by a single index.

Recently, much attention has been paid to the use of machine learning to analyze big data. Due to the success of image recognition using deep learning such as deep neural networks (NN), recurrent NN, convolutional NN, and so forth, the application of deep learning to tasks other than image recognition has been widely explored. Since internet exchange providers (IXPs) give the communication infrastructure to CDNs [17], another study [18] focuses on the traffic information of IXPs and applies machine learning by combining PeeringDB and RouteViews BGP data to detect 15 HGs. In the present study, we attempt to detect unknown HGs from AS-level topology data. First, we extract ASes from Tier 1 to Tier 3 from the AS relationship data published by CAIDA [19] and analyze them, including the features of complex networks. Next, we select a suitable machine learning method to classify the ASes. Using this method, we trained known CHs and detected unknown HGs. In this paper, Section II shows the AS features used in our study and III shows the method we used to analyze AS features as well as our method to detect HGs. The results of the analysis and detection are discussed in IV, and the features of the AS-level topology and the effectiveness of the HG detection are discussed in V.

II. AS Features

A. AS Relationship Features

CAIDA defines two main types of relationships between ASes: provider to customer (P2C) and peer to peer (P2P). In P2P links, ASes provide routing information and traffic to each other through nonmonetary contracts. This is called interconnection. In other cases, different ASes are merged into the same company's network, but the ASNs are retained. This is considered a sibling relationship. In this study, siblings are regarded as the same AS and are represented by the main AS, and P2C and P2P are the undirected relationships between two ASes constructing an AS-level topology, as shown in Fig. 1.

The traditional rendering of an AS topology model [20] is represented by a P2C hierarchical model with major ISPs at the top. These ASes that do not have any provider are called Tier 1. They are completely connected to each other by P2P links to form a creek. The CCS is a measure used to rank

TABLE I. AS FEATURES

AS relationship	Complex network
Provider counts	Degree centrality
Customer counts	Closeness centrality
Peer counts	Between centrality
Customer cone size	Eigenvector centrality
	Cluster coefficient

ASes defined by CAIDA, as mentioned above. The CCS of an AS is the total number of customer ASes that can be reached by P2C links plus one (for its own AS). Major ISPs, which are responsible for Internet connectivity, have large CCSes. In this study, the numbers of providers, customers, and peers, as well as their CCSes, are taken as the feature values of an AS relationship.

B. Network Topological Features

In complex network theory [21], various feature values are defined. In this study, we consider AS-level topology as a complex network and extract feature values. The four types of centrality and cluster coefficients defined in the theory of complex networks are defined in the AS-level topology as follows.

Degree centrality C_i^D is a feature that evaluates the number of links between AS_i and its neighbor ASes, defined in (1):

$$C_i^D = \frac{k_i}{N-1} \tag{1}$$

where, N is the total number of ASes, the denominator is the number of possible links, and k_i is the degree of AS_i, which is the number of links coming out of each AS. The larger C_i^D is, the more adjacent ASes AS_i has. The range is $0 \leq C_i^D \leq 1$.

Closeness centrality C_i^C is a feature that evaluates the closeness of AS_i to other ASes, defined in (2);

$$C_i^C = \frac{N-1}{\sum_{i,j \in G} d_{ij}} \tag{2}$$

where, G is the set of ASes and d_{ij} is the internode distance between AS_i and AS_j. The communication path between ASes is determined by AS path lists exchanged by ASes [22]. The AS path with the fewest ASes is selected from the AS path list in the BGP routing information. d_{ij} is the number of links in the path between AS_i and AS_j. The internode distance is obtained by analyzing the AS relationship data in this study. C_i^C is the inverse of the average internode distance between AS_i and all the other ASes. The larger C_i^C is, the smaller the average delay to the other ASes, although this relation is not strict because the topology is not router-level. The range is $0 \leq C_i^C \leq 1$.

Betweenness centrality C_i^B is a feature that represents the amount of involvement of AS_i in the path connecting two other nodes, defined in (3):

$$C_i^B = \frac{m_i}{{}_{N-1}C_2} \tag{3}$$

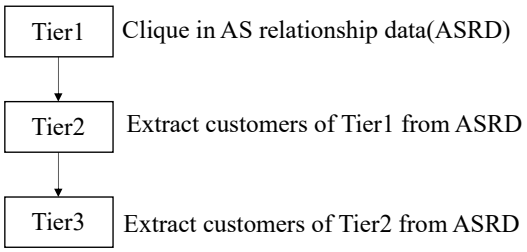


Fig. 2. Classification flow of ASes

where, m_i is the number of routes that include AS_i in the shortest path between both ASes except AS_i, and the denominator is the total number of combinations of two ASes except AS_i. The larger C_i^B is, the more important is its role in ensuring the connectivity of the communication channel. The range is $0 \leq C_i^B \leq 1$.

Eigenvector centrality C_i^{ev} is the first eigenvector corresponding to the eigenvalue with the largest absolute value, especially among the centralities reflecting those of neighboring ASes, using the eigenvectors of the adjacency matrix of the undirected graph, defined by (4);

$$C_i^{ev} = \frac{1}{\lambda} \sum_{\substack{j=1 \\ i \neq j}}^{N} a_{ij} C_j^{ev} \quad (4)$$

where, a_{ij} is the (i, j)-component of the adjacency matrix of AS-level topology and λ is the largest eigenvalue of the matrix. The larger C_i^{ev} is, the more important are the ASes that AS_i links. The range is $0 \leq C_i^{ev} \leq 1$.

A cluster is an agglomeration of nodes, and the smallest cluster is a triangle consisting of three nodes with three links. The cluster coefficient C_i is a measure of the degree of cluster formation involving AS_i, defined in (5);

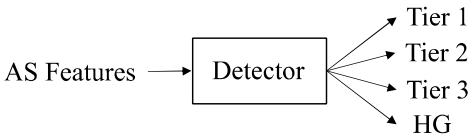
$$C_i = \frac{n_i}{{}_{k_i}C_2} \quad (5)$$

where, the denominator is the number of AS pairs adjacent to AS_i and n_i is the number of linked ones among the AS pairs. The larger C_i is, the more advanced is the cluster formation around AS_i. The AS relationship features and network topology features used in this study are listed in Table I.

III. DATA AND METHODS

A. Analysis of AS Features

In this study, we used the AS relationship data files published by CAIDA. Figure 2 shows the analysis flow of the AS relationship data. First, ASNs of the Tier 1 class were described in the first line of each AS relationship data file. Next, ASes that were not Tier 1 and appeared in P2C records containing a Tier 1 AS were classified as Tier 2. Then, ASes that were not yet classified and appeared in P2C records containing a Tier 2 AS were classified as Tier 3. When all ASes were classified in this way, the highest Tier class was Tier 7. However, the number of ASes in Tier 4 was much smaller than that in Tier 2 or Tier 3, and the number of links was also smaller. Meanwhile, when we analyzed the BGP path lists collected by the routers of stub ASes, we found that the actual number of ASes was larger than the number listed in the AS relationship data, and the links between subordinate ASes were dense [14]. Therefore, in this study we extracted and analyzed only Tier 3 or higher ASes and the links between them. Finally, 10 ASes were selected as known HGs, which were GAFAM and the major commonly known CHs and CDNs. Since all of them belonged to Tier 2 in the above classification, they were removed from the Tier 2 class. The details of the selected HGs are described in IV.



Train data : Tier1~Tier3, known HGs

Fig. 3. Detection of HGs by machine learning

TABLE II. TRAIN AND TEST DATA FOR MACHINE LEARNING

	Accuracy Test	Discovery (RF)
Train data	Tier 2 80% Tier 3 80%	Tier 1 100% (19 Ases) Tier 2 50% Tier 3 50% Known HG 100% (10 Ases)
Test data	Tier 2 20% Tier 3 20%	Tier 2 50% Tier 3 50%

Next we calculated the AS degree and CCS for each AS of Tier 1 to Tier 3 and for the known HGs. The AS degree was the sum of the number of provider, customer, and peer ASes. The CCS was obtained by tracing the P2C links among the ASes up to Tier 3. Most Tier 3 ASes had a CCS 1, but some had more because they had P2C links to Tier 2 ASes as providers. In other words, the P2C links of ASes did not necessarily result in a hierarchical structure with Tier 1 at the top. Therefore, in this study we defined CCS to be the value counted by limiting the P2C links from the AS to Tier 2. These features were calculated from the AS relationship data file using Python ver.3.6.12 [23].

The degree centrality, closeness centrality, betweenness centrality, eigenvector centrality, and cluster coefficients of the Tier 1 to Tier 3 ASes were calculated using NetworkX ver. 2.4 [24].

B. Detection of Unknown HGs

In this study, we propose a classifier of ASes and an HG detector from AS features by machine learning, as shown in Fig. 3. We examined the optimal machine learning algorithms for our purpose among the following machine learning algorithms. Random forest (RF) [25] collects several weak classifiers of decision trees and classifies something by ensemble learning. Support vector machines (SVM) [26] combined with linear SVM (SVM linear) and nonlinear SVM (SVM rbf) classify the nodes in the feature space by defining the boundary plane so that the distance between the node closest to the boundary plane and the boundary plane is maximized. The k-nearest neighbor method (k-KNN) [27] measures the Euclidean distance between pre-classified teacher nodes and unknown nodes in the feature space, takes k teacher nodes in order of proximity to the unknown nodes, and determines the classification of the unknown nodes by majority vote. Multilayer perceptron (MLP) [28] is a neural

TABLE III. KNOWN HGS

ASN	AS name	Counts				Type	Country
		CCS	Provider	Customer	P2P		
13335	CLOUDFLARENET	375	125	147	372	CDN	USA
20940	AKAMAI-ASN1	12	129	11	379	CDN	Netherlands
54113	FASTLY	1	33	0	259	CDN	USA
10310	YAHOO-1	38	11	37	182	eCommerce	USA
15169	GOOGLE	14	6	12	359	Search Engine	USA
8075	MICROSOFT-CORP-MSN-AS-BLOCK	9	14	7	283	Developer	USA
16509	AMAZON-02	5	32	4	312	eCommerce	USA
13414	TWITTER	4	9	3	254	SNS	USA
32934	FACEBOOK	4	14	3	376	SNS	USA
714	APPLE-ENGINEERING	2	23	1	292	Developer	USA

Apr 1, 2021

network with a structure of three or more layers. An MLP with four or more hidden layers is called a deep neural network (DNN). It performs supervised learning using the error back-propagation method.

These learning algorithms are known to be unsuitable for some adversaries, depending on the task. Therefore, to find out which algorithms are suitable as HG detectors, we performed Tier 2 and Tier 3 identification and compared the classification accuracies of machine learning algorithms. As shown in Table II, 80% of Tier 2 and Tier 3 ASes randomly selected and all Tier 1 ASes were training data. The remaining 20% of Tier 2 and Tier 3 ASes were test data. For machine learning tools, we used SVM, RF, and k-NN in Scikit-Learn ver. 0.23.1 [29]. k-NN was trained and tested with k=1, 3, 10, 100, and 1000. In addition, MLP (TensorFlow ver. 2.0.0 / Keras [30]) was used to train and test for 3, 5, and 7 intermediate layers. The number of nodes in the intermediate layers are 18/36/18, 18/36/72/36/18, and 18/36/72/144/72/36/18, respectively. Since the classification accuracy may vary depending on Tier 2 and Tier 3 ASes chosen for the training data, this procedure from data choice to classification was performed 10 times to obtain classification accuracy, and the average was taken. As a result, RF was used as the HG detector because RF had the highest accuracy as described in IV.

When detecting unknown HGs, we randomly divided Tier 2 and Tier 3 ASes into 50% each and designated them as Tier 2_A, Tier 3_A, Tier 2_B, and Tier 3_B, as shown in Table II. Tier 1, known HGs, Tier 2_A, and Tier 3_A were trained as training data, and Tier 2_B and Tier 3_B were used as test data to classify them into Tier 1, HG, Tier 2, and Tier 3. We also switched Tier 2_A and Tier 3_A with Tier 2_B and Tier 3_B and classified them again. The procedure from data selection to detection of A and B was performed 1000 times and the HGs were collected.

IV. RESULTS OF AS FEATURE ANALYSIS AND HG DETECTION

A. Analysis of AS Features

Classification results of the AS relationship data from 2019 to 2021 are shown in Table III. Many Tier 2 and Tier 3 ASes were connected to 19 Tier 1 ASes. Moreover, the numbers of Tier 2 and Tier 3 ASes increased. The correlation between the CCS published in CAIDA (CCSo) and that in this study (CCSp) was examined. For the data of July 1, 2021, CCSp = 1.1395 CCSo, and the coefficient of determination R^2 was 0.9683 for both Tier 1 and Tier 2. In contrast, CCSp = 0.4532 CCSo, and R^2 = 0.3843 for Tier 3. Therefore, the CCS correlation is strong for Tier 1 and Tier 2 but weak for Tier 3. The

TABLE IV. ASES CLASSIFIED BY TIER CLASS

AS class	Hypergiants	Tier 1	Tier 2	Tier 3	Total
Count	10	19	17436	35913	53368

Apr 1, 2019

AS class	Hypergiants	Tier 1	Tier 2	Tier 3	Total
Count	10	19	17956	38304	56289

Apr 1, 2020

AS class	Hypergiants	Tier 1	Tier 2	Tier 3	Total
Count	10	19	18487	41225	59731

Apr 1, 2021

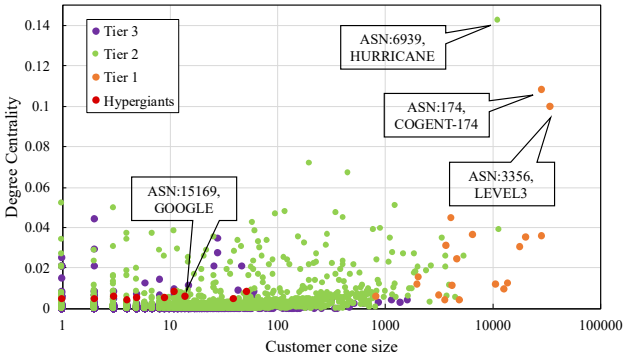


Fig. 4. Degree centrality vs Customer cone size

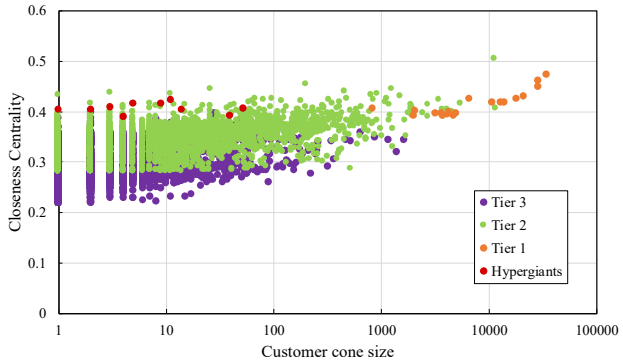


Fig. 5. Closeness centrality vs Customer cone size

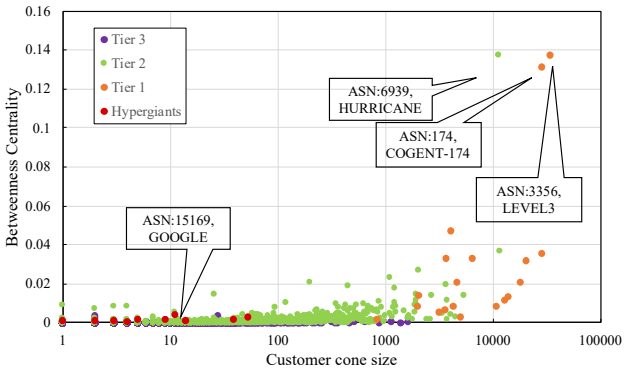


Fig. 6. Betweenness centrality vs Customer cone size

average values of CCSp were 12307.8, 17.2, and 1.82 for Tier 1, Tier 2, and Tier 3, respectively. Similarly, the average values of CCSo were 13852.1, 13.2, and 1.85, and the average numbers of P2P links that each AS had were 100.3, 18.5, and

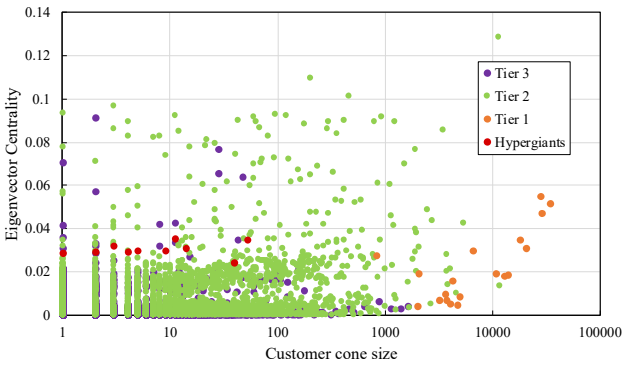


Fig. 7. Eigenvector centrality vs Customer cone size

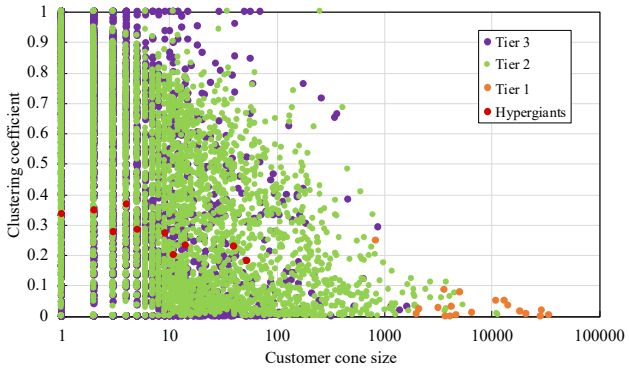


Fig. 8. Clustering coefficient vs Customer cone size

1.86 for Tier 1, Tier 2, and Tier 3, respectively. Ten HGs selected as known HGs are shown in Table IV. These HGs were classified as Tier 2 by P2C link analysis.

The results of the analysis of the complex network features are shown in Figs. 4 to 7. The horizontal axis of each figure is the logarithm of CCSp. As the figures show, Tier 1 ASes were distributed in areas with large CCSes, and Tier 2 ASes were distributed in wide areas. Although Tier 2 ASes have relatively larger CCSes than Tier 3 ASes, they appear to overlap in these figures. HGs are distributed in the area with relatively small CCSes.

The degree centrality of each AS against the CCSp is plotted in Fig. 4. Tier 1 and Tier 2 ASes tended to have large degree centralities. Especially, Level 3 (Tier 1), Cogent (Tier 1), and Hurricane (Tier 2) had extremely large degree centralities, indicating that the links were concentrated in these ASes. In contrast, HGs had small degree centralities. Closeness centralities were higher in Tier 1 and lower in Tier 3. HGs had high closeness centralities, as shown in Fig. 5. Betweenness centralities were very small for almost all ASes but extremely large for Level 3, Cogent, and Hurricane, as shown in Fig. 6. Almost eigenvector centralities were low, whereas some of Tier 2 ASes were distributed in higher areas, as shown in Fig. 7. Cluster coefficients of Tier 2 and Tier 3 ASes were widely distributed from 0 to 1, whereas those of Tier 1 ASes and HGs were low, as shown in Fig. 8. Finally, HGs had nearly the same values for all these centralities.

B. Detection of Unknown HGs

Table V shows the Tier 2 and Tier 3 classification results by RF, SVM (linear), SVM (rbf), k -NN, and MLP. The classification accuracy by RF was high, but the accuracies by both linear (linear) and nonlinear (rbf) of SVM were low. The accuracy by k-NN was higher than that by SVM but lower than that by RF. The accuracy by k-NN tended to decrease as

TABLE V. AS CLASSIFICATION ACCURACIES OF MACHINE LEARNING METHODS

RF	SVM				
0.982	linear			rbf	
	0.810			0.719	
	KNN				
K	1	3	10	100	1000
	0.933	0.933	0.922	0.882	0.821
	MLP				
Layer	9/18/36/18/2	9/18/36/72/36/18/2		9/18/36/72/144/72/36/18/2	
	0.919	0.904		0.885	

Apr 1, 2020

TABLE VI. DETECTED HGS IN 2019 ~ 2021 AS RELATIONSHIP DATA

ASN	AS name	Organization	Country
42	WOODYNET-1	WoodyNet	USA
2906	AS-SSI	Netflix Streaming Services Inc.	USA
15133	EDGECAST	MCI Communications Services, Inc. d/b/a Verizon Business	USA
36408	CDNETWORKSUS-02	CDNetworks Inc.	USA
44444	Forcepoint-Cloud-AS	Forcepoint Cloud Ltd	UK
46489	TWITCH	Twitch Interactive Inc.	USA
57976	BLIZZARD	Blizzard Entertainment, Inc	USA
199524	GCORE	G-Core Labs S.A.	Luxembourg

Apr 1, 2019

ASN	AS name	Organization	Country
2603	NORDUNET	NORDUnet	Denmark
2906	AS-SSI	Netflix Streaming Services Inc.	USA
14537	CL-1379-14537	Continent 8 LLC	USA
15133	EDGECAST	MCI Communications Services, Inc. d/b/a Verizon Business	USA
36351	SOFTLAYER	SoftLayer Technologies Inc.	USA
42473	AS-ANEXIA	ANEXIA Internetdienstleistungs GmbH	Austria
57976	BLIZZARD	Blizzard Entertainment, Inc	USA
199524	GCORE	G-Core Labs S.A.	Luxembourg

Apr 1, 2020

ASN	AS name	Organization	Country
14630	INVESCO	Invesco Group Services, Inc.	USA
14907	WIKIMEDIA	Wikimedia Foundation Inc.	USA
15133	EDGECAST	MCI Communications Services, Inc. d/b/a Verizon Business	USA
16276	OVH	OVH SAS	France
21859	ZNET	Zenlayer Inc	USA
57976	BLIZZARD	Blizzard Entertainment, Inc	USA
136907	HWCLOUDS-AS-AP	HUAWEI INTERNATIONAL PTE. LTD.	Singapore

Apr 1, 2021

the value of k increased. The classification accuracy by MLP was lower than that by RF. Furthermore, the greater the number of intermediate layers of MLP were added, the lower the identification accuracy became.

HGs in the AS relationship data on Apr 1, 2019, Apr 1, 2020, and Apr 1, 2021 were detected. As shown in Table VI, cloud services, CDNs, and ISPs, but also streaming video distribution services, game companies, and an investment company were extracted. These HGs had small CCSes but large numbers of peers, and their closeness centralities were high.

V. DISCUSSION

One of the features of known HGs is that they were classified as Tier 2 because they had direct links to Tier 1. On the other hand, the CCSes of known HGs were small. These Tier 2 known HGs also had an average of 1.6 P2C links with Tier 1 and in some cases had P2P links with Tier 1. In addition, there were many P2P links between Tier 2 and Tier 3

ASes. In terms of centrality, closeness centrality differed greatly by tier, with Tier 1 and HG having particularly high closeness centralities. The reason for the high closeness centrality of Tier 1 is that closeness centrality is higher when the path lengths from other ASes are smaller, and Tier 1 is located at the center of the AS topology. On the other hand, the large closeness centralities of HGs may be due to the increased number of links that favor access to HGs. Betweenness centralities are high only for two Tier 1 ASes and one Tier 2 AS were high, while those for the others are low. In the past, Tier 1 played an important role in ensuring connectivity, but as communication paths have been distributed due to the increase in links, the role of ensuring connectivity may be concentrated in certain ASes.

In the comparison of machine learning algorithms, neither linear SVM nor nonlinear SVM had high classification accuracy. Even for k-NN, the accuracy decreased when the value of k increased. These results suggest that it is difficult to discriminate between Tier 2 and Tier 3 based on the distance between features. However, the accuracy of MLP also decreased as the number of intermediate layers increased. On the other hand, RF had the highest classification accuracy, 98%. The HGs detected by the RF algorithm were characterized by small CCSes and large numbers of P2Ps, similar to the known HGs. The organizations that owned the detected ASes were examined, and fast-growing CH companies, such as game companies and content distribution networks, were found.

VI. Conclusions

The recent growth in both the size and number of ASes has affected the AS-level topology year by year. We have focused on its change and the knowledge obtained from AS-level topology. In this study, we analyzed the AS relationship data published by CAIDA to obtain the features of both AS relationships and complex networks. Based on the results, we applied to detect unknown HGs by machine learning method. As a result, currently growing ASes of CDNs, streaming video distribution services, and so forth were detected as HGs. In other words, we have shown that AS topology analysis and machine learning can be used to detect newly growing HGs. Information about such influential ASes will be useful for many IPSs that are trying to improve their customer service by preventing failures and formulating routing policies.

In the AS relationship data of CAIDA, however, information about lower-layer ASes is unreliable. In addition, it is difficult at the current situation to obtain information on traffic, which is important in order to understand communication networks. If the node and link information of lower-layer ASes and the traffic information of the entire Internet can be collected, the communication status will be clearer to every ISPs and the network operation of each AS will be more resilient to failures such as major disasters.

References

[1] J. Hawkinson and T. Bates, “Guidelines for creation, selection, and registration of an Autonomous System (AS),” Request for Comments: 1930 (Internet Engineering Task Force), 1996.

[2] IANA, ASN utilization count; https://www.iana.org/numbers/allocations/ . (2021.05.14)

[3] Cisco Annual Internet Report (018-2023) White paper, https://www.cisco.com/c/ja_jp/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html .(2021.05.14)

[4] CenturyLink (Level3 Communications); https://www.centurylink.com/business/resources/product-finder.html#networking .(2021.05.14)

[5] Internet Society; https://www.internetsociety.org/blog/2018/01/14000-incidents-2017-routing-security-year-review/ . (2021.05.14)

[6] CAIDA; https://www.caida.org/ . (2021.05.14)

[7] Looking Glass: https://lookinglass.org/ . (2021.05.14)

[8] RIPE ATLAS: https://atlas.ripe.net/about/ . (2021.05.14)

[9] E. Rapaport, I. Poese, P. Zilberman, O. Holschke, and R. Puzis, "Predicting traffic overflows on private peering," *arXiv preprint arXiv:2010.01380*, 2020.

[10] T. Böttger, G. Antichi, E. L.Fernandes, R. di Lallo, M.Bruyere, S Uhlig, and I. Castro, I., "The elusive internet flattening: 10 years of IXP growth," CoRR, 2018.

[11] E. Carisimo, C.Selmo, J. I. Alvarez-Hamelin, and A. Dhamdhere, " Studying the evolution of content providers in IPv4 and IPv6 internet cores," Computer Communications, no.145, pp. 54-65. 2019.

[12] T. Arnold, J. He, W.Jiang, M.Calder, I. Cunha, V. Giotsas, and E. Katz-Bassett, "Cloud provider connectivity in the flat internet," In Proceedings of the ACM Internet Measurement Conference, pp. 230-246, 2020.

[13] A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, , C. Dietzel, , et al., "A view of Internet Traffic Shifts at ISP and IXPs during the COVID-19 Pandemic, 2020. https://www.depositonce.tu-berlin.de/bitstream/11303/13209/4/feldmann_etal_2020.pdf

[14] A. Ishida, K. Endo, M. Teshi, and M. Harayama, “Contents Driven Change of AS-level Topology,” IEICE technical report, vol. 116, no. 65, pp.77-82, 2016 (in Japanese).

[15] M. Teshi and M. Harayama, “Data Distribution Index for Autonomous Systems,” IEICE Technical Report vol. 117, no. 386, CQ2017-95, p. 57-62, 2018 (in Japanese).

[16] H. Ido and M. Harayama, “AS-Level Topology Modeling focused on Peer Links,” IEICE Technical Report, vol. 118, no. 466, IN2018-87, p.19-24, 2019 (in Japanese).

[17] T.Böttger, F. Cuadrado, G. Tyson, I. Castro, and S. Uhlig, "A Hypergiant’s View of the Internet, "ACM SIGCOMM CCR, vol. 47, no.1, 2017.

[18] T. Böttger, F.Cuadrado, and S. Uhlig, "Looking for hypergiants in peeringDB," *ACM SIGCOMM Computer Communication Review*, vol. *48*, no.3, pp. 13-19. 2018.

[19] CAIDA, AS Relationship Data, http://data.caida.org/datasets/as-relationships/serial-1/ . (2021.05.14)

[20] G. Siganos, S.L. Tauro, and M. Faloutsos, “Jellyfish: A conceptual model for the as internet topology,” J. Communications and Networks vol.8, no.3, pp.339-350, 2006.

[21] F. A. Rodrigues, “Network centrality: an introduction. In A mathematical modeling approach from nonlinear dynamics to complex systems,” pp. 177-196, Springer, Cham, 2019.

[22] Y. Rekhter, T. Li, and S. Hares, “A Border Gateway Protocol 4 (BGP-4),” Request for Comments: 4271 (Internet Engineering Task Force), 2006.

[23] python, https://www.python.org/. (2021.05.14)

[24] networkX, https://networkx.org/. (2021.05.14)

[25] L. Breiman, “Random forests,” Machine learning, vol.45, no.1, pp.5-32, 2001.

[26] J. A. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” Neural processing letters, vol.9, no.3, pp.293-300, 1999.

[27] T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEE transactions on information theory, vol.13, no.1, pp.21-27, 1967.

[28] S. K. Pal and S. Mitra, “Multilayer perceptron, fuzzy sets, classification,”, IEEE transactions on neural networks, vol.3, no.5, pp.683-697, 1992.

[29] Scikit-Learn, https://scikit-learn.org/. (2021.05.14)

[30] Tensorflow, https://www.tensorflow.org/. (2021.05.14)